Risk Analysis, Vol. 19, No. 2, 1999

Combining Probability Distributions From Experts in Risk Analysis

Robert T. Clemen[1,2] and Robert L. Winkler[1]

This paper concerns the combination of experts' probability distributions in risk analysis, discussing a variety of combination methods and attempting to highlight the important conceptual and practical issues to be considered in designing a combination process in practice. The role of experts is important because their judgments can provide valuable information, particularly in view of the limited availability of ''hard data'' regarding many important uncertainties in risk analysis. Because uncertainties are represented in terms of probability distributions in probabilistic risk analysis (PRA), we consider expert information in terms of probability distributions. The motivation for the use of multiple experts is simply the desire to obtain as much information as possible. Combining experts' probability distributions summarizes the accumulated information for risk analysts and decision-makers. Procedures for combining probability distributions are often compartmentalized as mathematical aggregation methods or behavioral approaches, and we discuss both categories. However, an overall aggregation process could involve both mathematical and behavioral aspects, and no single process is best in all circumstances. An understanding of the pros and cons of different methods and the key issues to consider is valuable in the design of a combination process for a specific PRA. The output, a ''combined probability distribution,'' can ideally be viewed as representing a summary of the current state of expert opinion regarding the uncertainty of interest.

KEY WORDS: Combining probabilities; expert judgment; probability assessment.

1. INTRODUCTION

Expert judgments can provide useful information for forecasting, making decisions, and assessing risks. Such judgments have been used informally for many years. More formally, consulting several experts when considering forecasting or risk-assessment problems has become increasingly commonplace in the post-World War II era. Cooke (1991) reviews many of the developments over the years as attempts have been made to use expert judgments in various settings. Application areas have been diverse, including nuclear engineering, aerospace, various types of forecasting (economic, technological, meteorological, and snow avalanches, to name a few), military intelligence, seismic risk, and environmental risk from toxic chemicals.

In this paper we consider the problem of using multiple experts. Wu, Apostolakis, and Okrent (1990, p. 170) state that ''In PRA, an important issue related to knowledge representation under uncertainty is the resolution of conflicting information or opinions.'' Although we discuss judgments of various kinds, including forecasts, estimates, and probability assessments, our primary focus is on the aggregation and use of subjectively assessed probability distributions. The paper does not pretend to give a comprehensive view of the topic of group judgments; the accumu-

[1] Fuqua School of Business, Duke University, Durham, NC 27708-0120.
[2] To whom all correspondence should be addressed.

lated knowledge in this field springs from many disciplines, including statistics, psychology, economics, engineering, risk analysis, and decision theory. Our intent is to highlight the key issues involved in combining experts' probability distributions and to discuss a variety of combining methods.

Because the focus in this paper is on the *combination* of experts' probability distributions, we do not discuss the process by which such probability distributions might be elicited from individual experts. For general discussions of risk analysis, the reader is directed to basic sources (see Merkhofer (1987), Mosleh, Bier, & Apostolakis (1987), Bonano *et al*. (1990), Keeney & von Winterfeldt (1989, 1991), Morgan & Henrion (1990), Hora (1992), and Otway & von Winterfeldt (1992)). We believe that the elicitation protocol should be designed and conducted by a risk assessment team comprising individuals knowledgeable about both the substantive issues of interest and probability elicitation.

The mathematical and behavioral approaches we discuss in this paper assume that the experts have ironed out differences in definitions and that they all agree on exactly what is to be forecast or assessed. Practicing risk analysts know that these are strong assumptions and that much effort may be required in order to reach such agreement. Discussions of protocol development and use, such as those referenced in the previous paragraph, typically emphasize the importance of familiarizing the experts with the major issues in the problem at hand so that the experts will have a common understanding of the problem. Bonduelle (1987) specifically addresses the matter of resolving definitional disagreements among experts as an integral part of the aggregation problem.

Even though reasonable experts may agree on the definitions of relevant variables, it is, of course, still possible for them to disagree about probabilities for those variables. Such disagreement may arise for a multitude of reasons, ranging from different analytical methods to differing information sets or different philosophical approaches. Indeed, if they never disagreed there would be no point in consulting more than one expert. Morgan and Keith (1995, p. 468) note that the results of expert elicitations related to climate change "reveal a rich diversity of expert opinion." Consulting multiple experts may be viewed as a subjective version of increasing the sample size in an experiment. Because subjective information is often viewed as being "softer" than "hard scientific data," it seems particularly appropriate to consult multiple experts in an attempt to beef up the information base.

These motivations are reasonable; the fundamental principle that underlies the use of multiple experts is that a set of experts can provide more information than a single expert. Although it is sometimes reasonable to provide a decision maker with only the individual experts' probability distributions, the range of which can be studied using sensitivity analysis, it is often necessary to combine the distributions into a single one. In many cases, for example, a single distribution is needed for input into a larger model; if that model requires distributions for many variables, a full-blown sensitivity analysis may not be feasible.

Combination, or aggregation, procedures are often dichotomized into *mathematical* and *behavioral* approaches, although in practice aggregation might involve some aspects of each. Mathematical aggregation methods consist of processes or analytical models that operate on the individual probability distributions to produce a single "combined" probability distribution. For example, we might just take the averages of probabilities from multiple experts. Reviews of the literature on mathematical combination of probability distributions include Winkler (1968), French (1985), Genest and Zidek (1986), and Cooke (1991); Clemen (1989) reviews the broader area of combining forecasts (see also Bunn, 1988). Mathematical aggregation methods range from simple summary measures such as arithmetic or geometric means of probabilities to procedures based on axiomatic approaches or on various models of the information-aggregation process requiring inputs regarding characteristics such as the quality of and dependence among the experts' probabilities.

In contrast, behavioral aggregation approaches attempt to generate agreement among the experts by having them interact in some way. This interaction may be face-to-face or may involve exchanges of information without direct contact. Behavioral approaches consider the quality of individual expert judgments and dependence among such judgments implicitly rather than explicitly. As information is shared, it is anticipated that better arguments and information will be more important in influencing the group and that redundant information will be discounted.

In Sections 2 and 3, we discuss mathematical and behavioral methods, respectively, for combining experts' probability distributions. Some empirical results regarding these approaches are presented

in Section 4, and in Section 5 we summarize our views on the key issues in the combination of experts' probability distributions in risk analysis.

2. MATHEMATICAL COMBINATION METHODS

2.1. Axiomatic Approaches

Early work on mathematical aggregation of probabilities focused on axiom-based aggregation formulas. In these studies, the strategy was to postulate certain properties that the combined distribution should follow and then derive the functional form of the combined distribution. French (1985) and Genest and Zidek (1986) provide critical reviews of this literature, and our summary here draws heavily on these sources.

An appealing approach to the aggregation of probability distributions is the *linear opinion pool*, so named by Stone (1961), and dating back to Laplace (Bacharach, 1979):

$$p(\theta) = \sum_{i=1}^{n} w_i p_i(\theta), \tag{1}$$

where n is the number of experts, $p_i(\theta)$ represents expert i's probability distribution for unknown θ, $p(\theta)$ represents the combined probability distribution, and the weights w_i are non-negative and sum to one. For simplicity, we will use p to represent a mass function in the discrete case and a density function in the continuous case and will ignore minor technical issues involving the difference between the two cases in order to focus on the more important underlying conceptual and practical issues. As a result, we will often use "probabilities" as a shorthand for "probabilities or densities" or "probability distributions."

The linear opinion pool clearly is just a weighted linear combination of the experts' probabilities, and as such it is easily understood and calculated. Moreover, it satisfies a number of seemingly reasonable axioms. For example, it satisfies the *unanimity* property; if all of the experts agree on a probability, then the combined probability must also agree. Of particular note, the linear opinion pool is the only combination scheme that satisfies the *marginalization property* (MP). Suppose θ is a vector of uncertain quantities, and the decision maker is interested in just one element of the vector, θ_j. According to MP, the combined probability is the same whether one combines the experts' marginal distributions of θ_j or combines the experts' joint distributions of the vector θ and then calculates the marginal distribution of θ_j.

The weights in (1) clearly can be used to represent, in some sense, the relative quality of the different experts. In the simplest case, the experts are viewed as equivalent, and (1) becomes a simple arithmetic average. If some experts are viewed as "better" than others (in the sense of being more precise because of better information, for example), the "better" experts might be given greater weight. In some cases it is possible for some of the weights to be negative (Genest, 1984). The determination of the weights is a subjective matter, and numerous interpretations can be given to the weights (Genest & McConway, 1990).

Another typical combination approach uses multiplicative averaging and is sometimes called a *logarithmic opinion pool*. In this case, the combined probability distribution is of the form

$$p(\theta) = k \prod_{i=1}^{n} p_i(\theta)^{w_i}, \tag{2}$$

where k is a normalizing constant and the weights w_i satisfy some restrictions to assure that $p(\theta)$ is a probability distribution. Typically, the weights are restricted to sum to one. If the weights are equal (to $1/n$), then the combined distribution is proportional to the geometric mean of the individual distributions.

Formula (2) satisfies the principle of *external Bayesianity* (EB). Suppose a decision maker has consulted the experts, has calculated $p(\theta)$, but has subsequently learned some new information relevant to θ. Two choices are available. One is to use the information first to update the experts' probability distributions $p_i(\theta)$ and then combine them. The other is to use the information to update the combined $p(\theta)$ directly. A formula satisfies EB if the result is the same in each case.

Cooke (1991) presents a generalization of the linear and logarithmic opinion pools. This generalization begins by taking a weighted average not of the probability distributions, but of the probability distributions raised to the rth power. He then raises this weighted average to the $1/r$ power and normalizes it. When $r = 1$, this is the linear opinion pool, when r approaches zero it approaches the logarithmic opinion pool, and for other values of r it gives yet other combination rules.

Rules such as (1) or (2) may be quite reasonable, but not necessarily because of connections with properties such as MP or EB. Difficulties with the axioms

themselves are discussed elsewhere (French (1985) and Genest & Zidek (1986)). Lindley (1985) gives an example of the failure of both axioms in a straightforward example, with the interpretation that MP ignores important information and EB requires that the form of the pooling function not change. In addition, French (1985) points out that *impossibility theorems* exist (along the lines of Arrow's (1951) classic work on social choice theory) whereby a combining rule cannot satisfy simultaneously a number of seemingly compelling desiderata. Moreover, despite the work of Genest and McConway (1990) no foundationally based method for determining the weights in (1) or (2) is available.

2.2. Bayesian Approaches

French (1985), Lindley (1985), and Genest and Zidek (1986) all conclude that for the typical risk analysis situation, in which a group of experts must provide information for a decision maker, a Bayesian updating scheme is the most appropriate method. Winkler (1968) provides a Bayesian framework for thinking about the combination of information and ways to assess differential weights. Building on this framework, Morris (1974, 1977) formally establishes a clear Bayesian paradigm for aggregating information from experts. The notion is straightforward. If n experts provide information $g_1, \ldots, g_n$ to a decision-maker regarding some event or quantity of interest θ, then the decision maker should use Bayes' Theorem to update a prior distribution $p(\theta)$:

$$p^* = p(\theta \mid g_1, \ldots, g_n) \propto \frac{p(\theta)\, L(g_1, \ldots, g_n \mid \theta)}{p(g_1, \ldots, g_n)}, \quad (3)$$

where L represents the likelihood function associated with the experts' information. This general principle can be applied to the aggregation of any kind of information, ranging from the combination of point forecasts or estimates to the combination of individual probabilities and probability distributions. Resting on the solid ground of probability theory, including requirements of coherence as explicated by de Finetti (1937) and Savage (1954), Morris's Bayesian paradigm provides a compelling framework for constructing aggregation models. In the past two decades, attention has shifted from the axiomatic approach to the development of Bayesian combination models.

At the same time that it is compelling, the Bayesian approach is also frustratingly difficult to apply. The problem is the assessment of the likelihood function $L(g_1, \ldots, g_n \mid \theta)$. This function amounts to a probabilistic model for the information $g_1, \ldots, g_n$, and as such it must capture the interrelationships among θ and $g_1, \ldots, g_n$. In particular, it must account for the precision and bias of the individual g_is, and it must also be able to model dependence among the g_is. For example, in the case of a point forecast, the precision of g_i refers to the accuracy with which expert i forecasts θ and bias is the extent to which the forecast tends to fall consistently above or below θ. Dependence involves the extent to which the forecast errors for different experts are interrelated. For example, if expert i overestimates θ, will expert j tend to do the same?

The notions of bias, precision, and dependence are also crucial, but more subtle, in the case of combining probability distributions. Bias, for example, relates to the extent to which an expert can provide calibrated probability judgments. That is, an expert is empirically calibrated if, upon examining those events for which the expert judged an x percent chance of occurrence, it turns out that x percent; actually occur. Precision relates to the "certainty" of an expert; a calibrated expert who more often assesses probabilities close to zero or one is more precise. In the case of assessing a probability distribution for continuous θ, a more precise expert assessment is one that is both calibrated and at the same time has less spread (possibly measured as variance). Dependence among such distributions might refer to the tendency of the experts to report similar probabilities.

Because of the difficulty of assessing an appropriate likelihood function "from scratch," considerable effort has gone into the creation of "off-the-shelf" models for aggregating single probabilities (*e.g.,* Lindley, 1985; Clemen & Winkler, 1987) and probability distributions (Winkler, 1981; Lindley, 1983; Mendel & Sheridan, 1989). We will review a number of these models.

2.2.1. Bayesian Combinations of Probabilities

Suppose that θ is an indicator variable for a specific event, and the experts provide probabilities that $\theta = 1$ (i.e., that the event will occur). How should these probabilities be combined? Clemen and Winkler (1990) review and compare a number of different Bayesian models that might be applied in this situation. Here we review four of these models.

Let p_i ($i = 1, \ldots, n$) denote expert i's stated probability that θ occurs. Expressed in terms of the posterior odds of the occurrence of θ, $q^* = p^* / (1 - p^*)$, the models are as follows:

Independence:

$$q^* = \frac{p_0}{1 - p_0} \prod_{i=1}^{n} \frac{f_{1i}(p_i|q = 1)}{f_{0i}(p_i|q = 0)}, \tag{4}$$

where f_{1i} (f_{0i}) represents the probability of expert i giving probability p_i conditional on the occurrence (non-occurrence) of θ, and p_0 denotes the prior probability $p(\theta = 1)$. This model reflects the notion that each expert brings independent information to the problem of assessing p^*. Thus, more experts can mean more certainty. For example, if all experts say that the probability is 0.6, then p^* will tend to be much higher than 0.6.

Genest and Schervish:

$$q^* = \frac{p_0^{1-n} \prod_{i=1}^{n} p_0 + \lambda_i (p_i - \mu_i)}{(1 - p_0)^{1-n} \prod_{i=1}^{n} 1 - [p_0 + \lambda_i(p_i - \mu_i)]}, \tag{5}$$

where μ_i is the decision maker's marginal expected value of p_i and λ_i is interpreted as the coefficient of linear regression of θ on p_i. This model is from Genest and Schervish (1985) and is derived on the assumption that the decision maker (or assessment team) can assess only certain aspects of the marginal distribution of expert i's probability p_i. It is similar to the independence model, but allows for miscalibration of the p_is in a specific manner.

Bernoulli:

$$p^* = \sum_{i=1}^{n} \beta_i p_i. \tag{6}$$

This model, which is from Winkler (1968) and Morris (1983), invokes the idea that each expert's information is equivalent to a sample from a Bernoulli process with parameter θ. The resulting p^* is a convex combination of the p_is, with the coefficients interpreted as being directly proportional to the amount of information each expert has.

Normal:

$$q^* = \frac{p_0}{1 - p_0} \operatorname{Exp} [\mathbf{q}' \Sigma^{-1} (\mathbf{M_1} - \mathbf{M_0}) - (\mathbf{M}_1 + \mathbf{M}_0)' \Sigma^{-1} (\mathbf{M}_1 - \mathbf{M}_0)/2], \tag{7}$$

where $\mathbf{q}' = (\log[p_1/(1 - p_1)], \ldots, \log[p_n/(1 - p_n)])$ is the vector of log-odds given by the experts, a prime denotes transposition, and the likelihood functions for $\mathbf{q}$, conditional on $\theta = 1$ and $\theta = 0$, are modeled as normal with means $\mathbf{M}_1$ and $\mathbf{M}_0$, respectively, and common covariance matrix Σ. This model captures dependence among the experts' probabilities through the multivariate-normal likelihood functions, and is developed in French (1981) and Lindley (1985). Clemen and Winkler (1987) use this model in studying meteorological forecasts.

These four models all are consistent with the Bayesian paradigm, yet they are clearly all different. The point is not that one or another is more appropriate overall, but that different models may be appropriate in different situations, depending on the nature of the situation and an appropriate description of the experts' probabilities. Technically, these differences give rise to different likelihood functions, which in turn give rise to the different models.

2.2.2. Bayesian Models for Combining Probability Distributions

Just as the models above have been developed specifically for combining event probabilities, other Bayesian models have been developed for combining probability distributions for continuous θ. Here we review some of these models.

Winkler (1981) presents a model for combining expert probability distributions that are normal. Assume that each expert provides a normal distribution for θ with mean μ_i and variance σ^2_i. The vector of means $\mu = (\mu_1, \ldots, \mu_n)$ represents the experts' estimates of θ. Thus, we can work in terms of a vector of errors, $\varepsilon = (\mu_1 - \theta, \ldots, \mu_n - \theta)$. These errors are modeled as multivariate normally distributed with mean vector $(0, \ldots, 0)$ and covariance matrix Σ, regardless of the value of θ. Let $\boldsymbol{e}' = (1, \ldots, 1)$, a conformable vector of ones. Assuming a noninformative prior distribution for θ, the posterior distribution for θ is normal with mean μ^* and variance σ^{*2}, where

$$\mu^* = \boldsymbol{e}'\Sigma^{-1} \mu/\boldsymbol{e}' \Sigma^{-1}\boldsymbol{e} \tag{8a}$$

and

$$\sigma^{*2} = (\boldsymbol{e}' \Sigma^{-1}\boldsymbol{e})^{-1}. \tag{8b}$$

In this model the experts' stated variances σ_i^2 are not used directly (for an extension, see Lindley (1983)), although the decision-maker may let the ith diagonal element of Σ equal σ_i^2.

The normal model has been important in the development of practical ways to combine expert

judgments. The typical minimum-variance model for combining forecasts is consistent with the normal model (*e.g.,* see Bates & Granger, 1969; Newbold & Granger, 1974; Winkler & Makridakis, 1983). The multivariate-normal likelihood embodies the available information about the qualities of the experts' opinions, especially dependence among them. Biases can easily be included in the model via a non-zero mean vector for ε. Clemen and Winkler (1985) use this normal model to show how much information is lost owing to dependence, and they develop the idea of equivalent independent information sources. Winkler and Clemen (1992) show how sensitive the posterior distribution is when correlations are high, which is the rule rather than the exception in empirical studies. Chhibber and Apostolakis (1993) also conduct a sensitivity analysis and discuss the importance of dependence in the context of the normal model. Schmittlein, Kim, and Morrison (1990) develop procedures to decide whether to use weights based on the covariance matrix or to use equal weights. Similarly, Chandrasekharan, Moriarty, and Wright (1994) propose methods for investigating the stability of weights and deciding whether to eliminate some experts from the combination.

Although the normal model has been useful, it has some shortcomings. In particular, one must find a way to fit the expert judgments into the normal framework. If the experts have provided distributions that are unimodal and roughly symmetric, this is generally not a problem. Otherwise, some sort of transformation is required. The covariance matrix Σ typically is estimated from data; Winkler (1981) derives a formal Bayesian model when Σ is viewed as an uncertain parameter. Assessing Σ subjectively is possible; Gokhale and Press (1982) and Clemen and Reilly (1999) discuss the assessment of correlation coefficients via a number of different probability transformations. Finally, the posterior distribution is always a normal distribution, and typically is a compromise. For example, suppose two experts give $\mu_1 = 2$, $\mu_2 = 10$, and $\sigma_1^2 = \sigma_2^2 = 1$. Then the posterior distribution will be a normal distribution with mean $\mu^* = 6$ and variance $(1 + \rho)/2$, where ρ is the correlation coefficient between the two experts' errors. Thus, the posterior distribution puts almost all of the probability density in a region that neither of the individual experts thought likely at all. In a situation such as this (and assuming that the disagreement is *not* due to a disagreement about variable definition!), it might seem more reasonable to have a bimodal posterior distribution reflecting the two experts' opinions. Lindley (1983) shows how such a bimodal distribution can arise from a *t*-distribution model.

Another variant on the normal model is a Bayesian hierarchical model (Lipscomb, Parmigiani, & Hasselblad, 1998) that allows for differential random bias on the part of the experts by assuming that each expert's error mean in the normal model can be viewed as a random draw from a second-order distribution on the error means. Dependence among experts' probabilities arises through the common second-order distribution. Hierarchical models generally result in an effective shrinkage of individual means to the overall mean and tend to provide robust estimates. Because they involve another layer of uncertainty, they can be complex, particularly when the parameters are all viewed as unknown and requiring prior distributions.

Mendel and Sheridan (1989) develop a Bayesian model (also discussed in some detail in Cooke, 1991) that allows for the aggregation of probability distributions that are not necessarily normal. Assume that each expert always provides m fractiles of his or her distribution. For each expert, this defines $m + 1$ bins into which the actual outcome could fall. With n experts, an $(m + 1)^n$ array represents all of the possible joint outcomes that could conceivably occur. Each cell in this array represents the actual outcome falling into a specific bin for expert 1, a bin for expert 2, and so on for all n experts. Data on previous assessments and outcomes provides information about the likelihood of an outcome falling in each bin. The model incorporates this information in a Bayesian way, updating the distribution on the parameters representing the probability for each bin as new information is obtained.

When the experts provide a set of distributions for a new variable, θ, there are specific bins (at most $nm + 1$) into which the actual realization of θ may fall, owing to the way the experts' fractiles interleave. To find the aggregated probability distribution for θ, the distribution is normalized over the possible bins into which the actual θ may fall.

Mendel and Sheridan's model conveniently combines issues of both individual calibration and dependence, and it does so in capturing information about the occurrences of actual outcomes in the $(m + 1)^n$ array. In fact, this approach is perhaps most appropriately termed *joint calibration*, because it produces probability distributions that are based on a multivariate version of the traditional single-expert calibration approach. The notion of joint calibration is also discussed by Clemen (1986), who notes that Morris's

Bayesian aggregation paradigm can be viewed as a joint-calibration exercise. In addition, Clemen (1985) and Clemen and Murphy (1986) use a joint-calibration procedure similar in spirit to Mendel and Sheridan to study the combination of meteorological forecasts of precipitation.

Many of the methods above require the estimation of parameters of the likelihood function, and an often unspoken assumption is that past data will be used for such estimation. Although data-based methods might be viewed as an important part of the scientific process, it is often the case in risk analysis that past data are not available. In such situations, the decision-maker or the risk-assessment team can subjectively estimate the parameters (or assess prior distributions for the parameters). This can be a difficult task, however; consider the assessment of the elements of the covariance matrix in the normal model or of the probabilities associated with each cell in the $(m + 1)^n$ array in Mendel and Sheridan's model.

Some effort has been directed toward the development of Bayesian aggregation methods that are suitable for the use of subjective judgment in determining the likelihood function. Clemen and Winkler (1993), for example, present a process for subjectively combining point estimates from experts; the approach is based on the sequential assessment of conditional distributions among the experts' forecasts, where the conditioning is specified in an influence diagram. Formal attention has also been given to Bayesian models for situations in which the experts provide only partial specifications of their probability distributions (*e.g.,* moments, fractiles) or the decision maker is similarly unable to specify the likelihood function fully (Genest & Schervish, 1985; West, 1992; West & Crosse, 1992; Gelfand, Mallick, & Dey, 1995). Bunn (1975) develops a model that considers only which expert performs best on any given occasion and uses a Bayesian approach to update weights in a combination rule based on past performance.

Jouini and Clemen (1996) develop a method for aggregating experts' probability distributions in which the multivariate distribution (likelihood function) is expressed as a function of the marginal distributions. A *copula* function (*e.g.,* Dall'Aglio *et al.,* 1991) provides the connections, including all aspects of dependence, among the experts' judgments as represented by the marginal distributions. For example, suppose that expert i assesses a continuous density for θ, $f_i(\theta)$, with corresponding cumulative distribution function $F_i(\theta)$. Then Jouini and Clemen show that under reasonable conditions, the decision maker's posterior distribution is

$$P(\theta|f_1, \ldots, f_n) \propto c[1 - F_1(\theta), \ldots, 1 - F_n(\theta)] \prod_{i=1}^{n} f_i(\theta), \quad (9)$$

where c represents the copula density function.

In the copula approach, judgments about individual experts are entirely separate from judgments about dependence. Standard approaches for calibrating individual experts (either data-based or subjective) can be used and involve only the marginal distributions. On the other hand, judgments about dependence are made separately and encoded into the copula function. Regarding dependence, Jouini and Clemen suggest the use of a member of the Archimedian class of copulas, all of which treat the experts symmetrically in terms of dependence. If more flexibility is needed, the copula that underlies the multivariate normal distribution can be used (Clemen & Reilly, 1999).

In other work involving the specification of the likelihood function, Shlyakhter (1994) and Shlyakhter *et al.* (1994) develop a model for adjusting individual expert distributions to account for the well-known phenomenon of overconfidence, and they estimate the adjustment parameter for two different kinds of environmental risk variables. Hammitt and Shlyakhter (1999) show the implications of this model for combining probabilities.

In this section, we have discussed a number of mathematical methods for combining experts' probability distributions. A number of important issues should be kept in mind when comparing these approaches and choosing an approach for a given application. These issues include the type of information that is available (*e.g.,* whether full probability distributions are given by the experts or just some partial specifications of these distributions); the individuals performing the aggregation of probabilities (*e.g.,* the risk assessment team, a single decision-maker or analyst, or some other set of individuals); the degree of modeling to be undertaken (assessment of the likelihood function, consideration of the quality of the experts' judgments); the form of the combination rule (which could follow directly from modeling or could be taken as a primitive, such as a weighted average); the specification of parameters of the combination rule (*e.g.,* the weights); and the consideration of simple vs. complex rules (*e.g.,* simple averages versus more complex models). The empirical results

in Section 4 will shed some light on some of these questions, and we will discuss the issues further in Section 6.

3. BEHAVIORAL APPROACHES

Behavioral combination approaches require experts to interact in some fashion. Some possibilities include face-to-face group meetings, interaction by computer, or sharing of information in other ways. The group may assess probabilities or forecasts, or simply discuss relevant issues and ideas with only informal judgmental assessment. Emphasis is sometimes placed on attempting to reach agreement, or consensus, within the group of experts; at other times it is simply placed on sharing of information and having the experts learn from each other. The degree to which the risk-assessment team is involved in structuring and facilitating the expert interaction can vary. Specific procedures (*e.g.,* Delphi, Nominal Group Technique) can be used, or the risk-assessment team can design the interaction process to suit a particular application.

The simplest behavioral approach is to assemble the experts and assign them the task of generating a ''group'' probability distribution. Discussion and debate can be wide-ranging, presumably resulting in a thorough sharing of individual information. Ideally, such sharing of information leads to a consensus, and under certain circumstances agreement ''should'' be reached (*e.g.,* Aumann, 1976). In practice, though, experts often do not agree. Any group probabilities that arise from the discussion may require the experts to negotiate and compromise in some way. In this sense, the so-called consensus may not reflect any one expert's thoughts perfectly. Instead, it is a statement that each expert is willing to accept for the sake of the group appearing to speak with one voice.

Group interaction can suffer from many problems (*e.g.,* see Janis & Mann, 1977). Some individuals may tend to dominate the discussions, new ideas may be discouraged, or the group may ignore important information. The phenomenon of group polarization may occur, in which a group tends to adopt a position more extreme than its information warrants (Plous, 1993). Hogarth (1977) points out, however, that interaction processes need not be dysfunctional. Having experienced analysts serve as facilitators for the group can improve the process. For example, decision conferencing (Phillips, 1984, 1987) is a particular approach to structuring group discussion for decision-making and group probability assessment. The role of facilitators is very important in decision conferencing, and Phillips and Phillips (1990) claim that the main role of a facilitator should be to control the process and structure of the group interaction (thereby trying to head off dysfunctional aspects of interaction), not to contribute substantively to the content of the discussion.

One of the oldest approaches to structuring group judgments, the Delphi method, requires indirect interaction (Dalkey, 1969; Linstone & Turoff, 1975; Parenté & Anderson-Parenté, 1987). Although different variations exist, experts typically make individual judgments (in our case, assessments of probability distributions), after which judgments are shared anonymously. Each expert may then revise his or her probabilities, and the process may be repeated. Ideally, the experts would agree after a few rounds, but this rarely happens. At the end of the Delphi rounds, the experts' final probability distributions typically are combined mathematically.

The Nominal Group Technique (Delbecq, Van de Ven, & Gustafson, 1975) is a related behavioral aggregation approach. Experts first assess their probability distributions individually and then present those distributions to other group members. Group discussion follows with the assistance of a facilitator, after which the experts may revise their probabilities. As with Delphi, the final probability distributions may require mathematical aggregation. Lock (1987) proposes a similar behavioral approach that stresses careful task definition and advocacy of alternative perspectives.

A recently proposed aggregation method described by Kaplan (1990) is designed to account explicitly for the information available to the group. In Kaplan's approach, a facilitator/analyst first leads the experts through a discussion of the available information. The objective of this discussion is to determine the ''consensus body of evidence'' for the variable of interest. When consensus is reached regarding the relevant information, the analyst proposes a probability distribution, conditioned on the consensus body of evidence. At this point, the analyst must obtain assurance from the experts that the evidence has been interpreted correctly in arriving at the probability distribution. Because different experts may interpret the evidence in different ways, the ''group judgment'' may differ from the individual experts' judgments and may result from something like a negotiation process among the experts.

Our discussion in this section has been brief, not

because behavioral combination approaches are not important but because they are, for the most part, not a set of fixed procedures. Delphi, the Nominal Group Technique, decision conferencing, and Kaplan's approach are exceptions. There are, however, a number of important issues relating to behavioral combination: the type of interaction (*e.g.,* face-to-face, via computer, anonymous); the nature of the interaction (*e.g.,* sharing information on relevant issues, sharing probabilities, trying to assess ''group probabilities''); the possibility of individual reassessment after interaction; and the role of the assessment team (*e.g.,* as facilitators). Although the discussion of behavioral approaches is brief, the empirical evidence regarding behavioral combination and related topics such as group decision making is more extensive, as we shall see in Section 4.

4. EMPIRICAL EVIDENCE

The various combining techniques discussed in Sections 2 and 3 all have some intuitive appeal, and some have a strong theoretical basis given that certain assumptions are satisfied. The proof, of course, is in the pudding. How do the methods perform in practice? Do the combination methods lead to ''improved'' probability distributions? Do some appear to perform better than others? Some evidence available from experimentation, analysis of various data sets, and actual applications, including work on combining forecasts that is relevant to combining probabilities, is reviewed in this section.

4.1. Mathematical Versus Intuitive Aggregation

Before comparing different combination methods, we should step back and ask whether one should bother with formal aggregation methods. Perhaps it suffices for the decision-maker or the risk-assessment team to look at the individual experts' probability distributions and to aggregate them intuitively, directly assessing a probability distribution in light of the experts' information.

Hogarth (1987, Ch. 3) discusses the difficulty humans have in combining information from different data sources. Although his discussion covers the use of all kinds of information, his arguments apply to the aggregation of expert opinion. Among other phenomena, Hogarth shows how individuals tend to ignore dependence among information sources, and he relates this to Kahneman and Tversky's (1972) ''representativeness'' heuristic. In a broad sense, Hogarth's discussion is supported by psychological experimentation showing that expert judgments tend to be less accurate than statistical models based on criteria that the experts themselves claim to use. Dawes, Faust, and Meehl (1989) provide a review of this literature.

Clemen, Jones, and Winkler (1996) study the aggregation of point forecasts. They use Winkler's (1981) normal model and Clemen and Winkler's (1993) conditional-distributions model, comparing the probability distributions derived from these models with intuitively assessed probability distributions. Although their sample size is small, the results suggest that the mathematical methods perform somewhat better than intuitive assessment, and the authors speculate that this is due to the structured nature of the assessments required in the mathematical-aggregation models.

4.2. Comparisons Among Mathematical Methods

Some evidence is available regarding the relative performance of various mathematical aggregation methods. In an early study, Staël von Holstein (1972) studied averages of probabilities relating to stock market prices. Most of the averages performed similarly, with weights based on rankings of past performance slightly better than the rest.

Seaver (1978) evaluated simple and weighted averages of individual probabilities. The performance of the different combining methods was similar, and Seaver's conclusion was that simple combination procedures, such as an equally weighted average, produces combined probabilities that perform as well as those from more complex aggregation models. Ferrell (1985) reached the same conclusion in his review of mathematical aggregation methods, and Clemen and Winkler (1987) reported similar results in aggregating precipitation probability forecasts.

In a follow-up study, Clemen and Winkler (1990) studied the combination of precipitation forecasts using a wider variety of mathematical methods. One of the more complex methods that was able to account for dependence among the forecasts performed best. Although a simple average was not explicitly considered, a weighted average that resulted in weights for the two forecasts that were not widely different performed almost as well as the more complex scheme.

Winkler and Poses (1993) report on the combination of experts' probabilities in a medical setting. For each patient in an intensive care unit, four individuals (an intern, a critical care fellow, a critical care attending, and a primary attending physician) assessed probabilities of survival. All possible combinations (simple averages) of these four probabilities were evaluated. The best combination turned out to be an average of probabilities from the two physicians who were simultaneously the most experienced and the least similar, with one being an expert in critical care and the other having the most knowledge about the individual patient.

All of these results are consistent with the general message that has been derived from the vast empirical literature on the combination of point forecasts. That message is that, in general, simpler aggregation methods perform better than more complex methods. Clemen (1989) discusses this literature. In some of these studies, taking into account the quality of the expert information, especially regarding relative precision of forecasts, turns out to be valuable.

The above studies focus on the combination of point forecasts or event probabilities, and mathematical methods studied have been either averages or something more complex in which combination weights were based on past data. Does the result that simpler methods work better than more complex methods carry over to the aggregation of probability distributions, especially when the quality of the expert opinions must be judged subjectively? Little specific evidence appears to be available on this topic. Clemen, Jones, and Winkler (1996) reported that Winkler's (1981) normal model and the more complex conditional-distributions model (Clemen & Winkler, 1993) performed at about the same level.

Even though little evidence exists on the performance of combination methods for probability distributions, do other kinds of evidence exist that might shed light on the performance of the different kinds of methods? The answer is a qualified ''yes.'' The Bayesian approach to aggregating expert probabilities can be viewed as an exercise in decomposition. Rather than a decision maker holistically assessing a probability distribution in light of expert reports, the Bayesian approach decomposes the problem into one of making assessments about expert judgments that are then recomposed via Bayes Theorem. Thus, for prospective evidence on the performance of Bayesian approaches, we appeal to the growing literature on the value of decomposing probability judgments into smaller and more manageable assessment tasks. Ravinder, Kleinmuntz, and Dyer (1988) provide a theoretical argument for the superiority of decomposed assessments. Wright, Saunders, and Ayton (1988), though, found little difference between holistic and decomposed probability assessments; on the other hand, Hora, Dodd, and Hora (1993) provide empirical support for decomposition in probability assessment. With regard to decomposition and the assessment of point estimates, Armstrong, Denniston, and Gordon (1975) and MacGregor, Lichtenstein, and Slovic (1988) found that decomposition was valuable in improving the accuracy of those estimates. Morgan and Henrion (1990) review the empirical support for decomposition in probability assessment, and Bunn and Wright (1991) do the same for forecasting tasks in general. A tentative conclusion is that, for situations in which the aggregation must be made on the basis of subjective judgments, appropriate decomposition of those judgments into reasonable tasks may lead to better performance. Moreover, decomposition enables the use of different sets of experts for different variables, thereby matching expertise to assessment task.

4.3. Behavioral Approaches

Perhaps the best known results from the behavioral group–judgment literature is group polarization, the tendency of groups to adopt more extreme positions than would individual members. ''Groupthink'' (Janis, 1982) is an extreme example of this phenomenon. According to Plous (1993), hundreds of studies of group polarization have been performed over the years, with the consistent conclusion that after discussion, a group will typically advocate a riskier course of action than they would if acting individually or without discussion.

The results on group polarization would appear to suggest caution when using behavioral combination methods. However, it is important to realize that the results on group polarization apply primarily to unstructured group discussions. It has been shown that group polarization can be deterred by such measures as delaying commitment of the group, spreading power among members, seeking additional information, and encouraging conflict among members (see Park, 1990). In the context of group probabilities, the polarization results indicate that groups might tend to be overconfident about their conclusions. Sniezek (1992) reviews work on confidence assessment in group decision-making and concludes that groups are

more confident than individuals and that they appear to be overconfident. In an experimental study of group decision-making behavior, Innami (1994) finds that the quality of group decisions increases to the extent that group members exchange facts and reasons (a ''reasoning'' orientation) and decreases to the extent that group members stick to their positions (a ''positional'' orientation), and that an intervention that emphasizes a knowledge-based logical discussion and consensual resolution of conflicts improves the quality of group decisions.

Experimental conclusions with respect to group judgment have been mixed. For example, a few studies on group-level judgment suggest caution. Myers and Lamm (1975) report evidence that face-to-face interaction in groups working on probability judgments may lead to social pressures that are unrelated to group members' knowledge and abilities. Gustafson *et al.* (1973), Fischer (1975), Gough (1975), and Seaver (1978) found in their experiments that interaction of any kind among experts led to increased overconfidence and hence poorer calibration of group probability judgments. More recently, Argote, Seabright, and Dyer (1986) found that the representativeness heuristic is used more by groups than by individuals, presumably leading to more biases (*e.g.,* overconfidence) related to this heuristic. In a related study, Tindale, Sheffey, and Filkins (1990) concluded that groups committed the conjunction fallacy more often than individuals. All of these results are generally consistent with the notion of group polarization.

A number of studies have examined the accuracy of group judgments. In a review, Hastie (1986) considers quantity estimation (comparable to point forecasting), problem solving, and answering almanac questions, and Gigone and Hastie (1997) update this review by considering relevant work that has appeared in the intervening time. In general, groups tend to perform better than the average individual, but the best individual in a group often outperforms the group as a whole. Looking only at quantity estimation (most pertinent for risk-assessment studies), the conclusion is that groups are only slightly (one-eighth of a standard deviation) more accurate than individuals on average. More recently, Sniezek and Henry (1989, 1990) have produced experimental evidence that the group's advantage in quantitative estimation may be somewhat greater than was reported by Hastie.

Another review by Hill (1982), examining 50 years of research on decision-making, reported similar results. She found that group performance is typically better than the average group member, but not as good as the best member. Einhorn, Hogarth, and Klempner (1977) and Uecker (1982) report experiments in which groups performed at the level of the best individual member. Both cases used students as subjects, and both allowed unlimited exchange of information but required that the group reach a consensus.

A related finding is that group accuracy often depends on the rules used by the group to arrive at a single judgment. Sniezek (1989) compared five types of group-aggregation methods. Four were behavioral-aggregation methods, and the fifth was a simple average of individual judgments. All four of the behavioral methods were more accurate than the average, but of those four, the best results by far were obtained by the ''dictator'' rule, in which the group selects on the basis of discussion a single individual whose opinion the group adopts. In an interesting twist, though, the chosen spokesperson always modified his or her opinion to be closer to the group average, thereby decreasing group accuracy slightly. In this case, the subjects again were college students, and the task was sales forecasting.

One of Sniezek's five group techniques was Delphi, and her results were similar to earlier studies on the accuracy of Delphi (*e.g.,* Dalkey, 1969; Dalkey & Brown, 1971). These studies showed that forecasts of individuals converged and that the Delphi technique performed slightly better than a similar procedure with face-to-face open interaction groups. In a study of bankers, Brockhoff (1975) found the same results for almanac questions, but face-to-face interaction provided better forecasts for economic forecasting. More recent work on Delphi has led to mixed results; Hastie (1986) and Parenté and Anderson-Parenté (1987) review this literature. The conclusion appears to be that Delphi has no clear advantage over other behavioral combination methods.

4.4. Mathematical Versus Behavioral Aggregation

Most of the research comparing mathematical and behavioral aggregation has focused on comparisons with a simple average of forecasts or probabilities rather than with more complicated mathematical combination methods. Results from these comparisons have been mixed. For example, for forecasting college students' grade-point averages, Rohrbaugh (1979) found that behavioral aggregation worked better than taking simple averages of individual group

members' forecasts. Hastie (1986), Hill (1982), and Sniezek (1989) reached similar conclusions, as described above. However, Lawrence, Edmundson, and O'Connor (1986) reported that mathematical combination improved on the behavioral combination of forecasts. In Flores and White's (1989) experiment, mathematical and behavioral combinations performed at approximately the same level. Goodman (1972) asked college students to assess likelihood ratios in groups and individually; the behavioral combination showed slight improvement over the mechanical combination.

Seaver (1978) asked student subjects to assess discrete and continuous probability distributions for almanac questions. Several different conditions were used: individual assessment, Delphi, Nominal Group Technique, free-form discussion, and two other approaches that structured information sharing and discussion. Both simple and weighted averages of the individual probabilities were also calculated. The conclusion was that interaction among the assessors did not improve on the performance of the aggregated probabilities, although the subjects did feel more satisfied with the behavioral aggregation results.

Reagan-Cirincione (1994) used an intensive group process intervention involving cognitive feedback and a computerized group support system for a quantity estimation task. The results show this to be the only study reviewed by Gigone and Hastie (1997) for which group judgments are more accurate than a mathematical average of the individual judgments. In general, Gigone and Hastie (1997) conclude that the evidence indicates that a simple average of individual judgments tends to outperform group judgments. Moreover, they discuss ways in which groups might improve over mathematical combinations and conclude that "there is a limited collection of judgment tasks in which groups have a legitimate opportunity to outperform individual judgments" (p. 162).

5. CONCLUSION

We have reviewed a variety of methods for combining probability distributions in risk analysis. The empirical results reviewed in Section 4 suggest that mathematical aggregation outperforms intuitive aggregation and that mathematical and behavioral approaches tend to be similar in performance, with mathematical rules having a slight edge. A comparison of behavioral approaches yields no clear-cut conclusions. As for mathematical combination methods, simple combination rules (*e.g.,* a simple average) tend to perform quite well. More complex rules sometimes outperform the simple rules, but they can be somewhat sensitive, leading to poor performance in some instances. All of these conclusions should be qualified by noting that they represent tendencies over a series of different empirical studies, generally conducted in an experimental setting as opposed to occurring in the context of a real-world risk analysis. These studies do not, unfortunately, directly assess the precise issue that needs to be addressed. For the purpose of the typical risk analysis in which probability distributions are to be combined, but limited past data are available, the real question is, "What is the best way to combine the judgments?" Thus, although we should pay careful attention to available empirical results and learn from them, we should think hard about their generalizability to realistic risk-analysis applications.

Both the mathematical combination of probabilities with some modeling and the use of interaction among the experts have some intuitive appeal. It is somewhat disappointing, therefore, to see that modeling and behavioral approaches often provide results inferior to simple combination rules. We feel a bit like the investor who would like to believe that some careful analysis of the stock market and some tips from the "pros" should lead to high returns but finds that buying a mutual fund which just represents a stock market index such as the S & P 500 would yield better returns. On the other hand, we should remember that the simple combination rules do perform quite well, indicating that the use of multiple experts and the combination of probabilities from these experts can be beneficial. One message that comes from the work on the combination of probabilities is that, at a minimum, it is worthwhile to consult multiple experts and combine their probabilities.

Another message is that further work is needed on the development and evaluation of combination methods. The challenge is to find modeling procedures or behavioral approaches (or processes involving both modeling aspects and behavioral aspects) that perform well enough to justify the extra cost and effort that is associated with serious modeling or expert interaction. On the behavioral side, Davis (1992, p. 34) states: "The engineering of increases in decision performance while maintaining [the advantages of group decision making] is a proper challenge for fundamental theory and research in applied psychology." Gigone and Hastie (1997, p. 166) echo this

in their concluding comments: ''The quality of group decisions and judgments is an essential ingredient in democratic institutions and societies, but research on group judgment accuracy is stagnant. . . . Better methods and analyses will help behavioral scientists, engineers, and policymakers to design and select group decision-making procedures that will increase efficiency, justice, and social welfare.''

Regarding mathematical combining procedures, we believe that simple rules will always play an important role, because of their ease of use, robust performance, and defensibility in public policy settings where judgments about the quality of different experts are eschewed. Simple rules do not, however, allow explicit consideration of factors such as overconfidence and dependence among experts. A modeling challenge is to develop mathematical approaches that facilitate modeling and assessment, are robust (*e.g.,* avoid extreme situations such as highly negative weights), and lead to improved performance. Chhibber, Apostolakis, and Okrent (1994, p. 102) write as follows: ''The Bayesian aggregation tool is demonstrably powerful, but it is not well understood . . . Further studies to understand its behavior . . . need to be undertaken to realize its full potential.'' In principle, the Bayesian approach allows for careful control in adjusting for the quality of expert distributions, including overconfidence and dependence. It is also worth noting that the normal and copula models permit the symmetric treatment of the experts, in which case simple combining rules fall out of these models as special cases.

More generally, the process of combining probability distributions in risk analysis may well involve both mathematical and behavioral aspects and should be considered in the context of the overall process for obtaining and utilizing expert judgment in a given application (for discussions of this process, see the references in Section 1). Important issues to consider include the following.

- *Selection of experts*. Having dealt with the combination of probabilities from experts in this paper, we see natural implications for the selection of experts whose probabilities are to be combined. Experts who are very similar (in modeling style, philosophy, access to data, etc.) tend to provide redundant information, and the high level of dependence means not only minimal gains from aggregation but also difficulties with some modeling approaches due to multicollinearity and the extreme (very large positive or negative) weights that can result. Thus, heterogeneity among experts is highly desirable. In terms of the number of experts, Makridakis and Winkler (1983) and Clemen and Winkler (1985) demonstrate the diminishing marginal returns associated with large numbers of experts. Their analyses, further supported by Ferrell (1985), suggest using three to five experts.
- *Flexibility and process design*. We believe that there is no single, all-purpose combining rule or combining process that should be used in all situations. Rather, the design of the combining process (as part of the overall expert judgment process) should depend on the details of each individual situation. This process design, conducted by the decision-maker (or decision-making body) in conjunction with the risk assessment team, should take into account factors such as the nature and importance of the uncertainties, the availability of appropriate experts, past evidence available about the experts and about the quantities of interest, the degree of uncertainty about quantities of interest, the degree of disagreement among the experts, the costs of bringing experts together, and a variety of other factors. We believe that a carefully structured and documented process is appropriate.
- *The role of modeling versus rules.* Risk analysis applications involving combination have often used simple combination rules, usually a simple average. Such simple rules are valuable benchmarks, but careful consideration should be given to modeling in order to include, in a formal fashion, factors such as the quality of the judgments from individual experts and the dependence among experts. One possible scenario is that the experts are judged to be exchangeable and their probabilities should be treated symmetrically, but this should be a conscious choice on the part of the risk assessment team. The degree of modeling will vary from case to case, ranging from fairly simple modeling (*e.g.,* unequal weights based on judgments of relative precision) to more detailed modeling (*e.g.,* building a full copula model). When little information is available about the relative quality of and dependence among the experts' probabilities, a simple rule such as a simple average is recommended.
- *The role of interaction*. This aspect of the com-

bination process will also vary from case to case, depending on such factors as the perceived desirability of exchanging information and the ease with which such information can be exchanged. Evidence to date does not provide strong support for benefits of interaction in the actual aggregation process, yet it has considerable intuitive appeal and has been used in risk analysis applications (*e.g.,* EPRI, 1986; Hora & Iman, 1989; Winkler *et al.,* 1995). We feel that the jury is still out on the impact of interaction on the quality of the resulting combined probabilities and that any benefits are most likely to come from exchanges of information (possibly including individual experts' probability distributions) as opposed to forced consensus through group probability assessments. This implies that mathematical combination will still be needed after interaction and individual probability assessment or reassessment. Also, it is important that the interaction process by carefully structured with extensive facilitation (preferably by the risk assessment team). We emphasize that our conclusions here relate specifically to the aggregation process; we believe that interaction is valuable in ironing out differences in definitions and assumptions, clarifying what is to be forecast or assessed, and exchanging information.

- *The role of sensitivity analysis*. It is helpful to conduct a sensitivity analysis to investigate the variation in the combined probabilities as parameters of combining models are varied. This can help in decisions regarding the scope of the modeling effort. A related note is that reporting the individual experts' probabilities as well as any combined probabilities provides useful information about the range of opinions in a given case as well as about the likely sensitivity of combined probabilities to different combining procedures.
- *Role of the risk-assessment team*. As should be clear from the above discussion, the risk-assessment team plays a very important role in the combination of probability distributions as well as in all other aspects of the expert judgment process. With respect to mathematical combination, the team should perform any modeling, make any assessments of expert quality that are needed in the modeling process, and choose the combination rule(s) to be used. On the behavioral side, they should structure any interaction and serve as facilitators. In general, they are responsible for the design, elicitation, and analysis aspects of the combination process.

In summary, the combination of experts' probability distributions in risk analysis is valuable for encapsulating the accumulated information for risk analysts and decision-makers and providing the current state of expert opinion regarding important uncertainties. Normatively and empirically, combining can lead to improvements in the quality of probabilities. More research is needed on the potential benefits of different modeling approaches and the development of mathematical combination rules. Likewise, continued research can lead to a better understanding of the cognitive and social psychology of group judgments, with the goal of further developing useful behavioral aggregation procedures. The ability to use wisely in practice both mathematical and behavioral aggregation methods, whether separately or in tandem, can contribute greatly to the practice of risk analysis.

ACKNOWLEDGMENTS

This research was supported in part by the National Science Foundation under Grants SBR-93-20754, SBR-93-20662, SBR-94-22527, and SBR-95-96176.

REFERENCES

Argote, L., Seabright, M. A., & Dyer, L. (1986). Individual versus group use of base-rate and individuating information. *Organizational Behavior and Human Decision Processes,* **38,** 65–75.

Armstrong, J. S., Denniston, W. B., & Gordon, M. M. (1975). The use of the decomposition principle in making judgments. *Organizational Behavior and Human Performance,* **14,** 257–263.

Arrow, K. J. (1951). *Social Choice and Individual Values.* New York: Wiley.

Aumann, R. J. (1976). Agreeing to Disagree. *Annals of Statistics,* **4,** 1236–1239.

Bacharach, M. (1979). Normal Bayesian dialogues. *Journal of the American Statistical Association,* **74,** 837–846.

Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. *Operational Research Quarterly,* **20,** 451–468.

Bonano, E. J., Hora, S. C., Keeney, R. L., & von Winterfeldt, D. (1990). *Elicitation and Use of Expert Judgment in Performance Assessment for High-Level Radioactive Waste Repositories.* NUREG/CR-5411. Washington, DC: Nuclear Regulatory Commission.

Bonduelle, Y. (1987). *Aggregating Expert Opinions by Resolving Sources of Disagreement.* PhD Dissertation, Stanford University.

Brockhoff, K. (1975). The performance of forecasting groups in computer dialogue and face-to-face discussion. In H. Linstone & M. Turoff (Eds.), *The Delphi Method: Techniques and Applications.* Reading, MA: Addison-Wesley.

Bunn, D. W. (1975). A Bayesian approach to the linear combination of forecasts. *Operational Research Quarterly,* **26,** 325–329.

Bunn, D. W. (1988). Combining forecasts. *European Journal of Operational Research,* **33,** 223–229.

Bunn, D., & Wright, G. (1991). Interaction of judgmental and statistical forecasting methods: Issues and Analysis. *Management Science,* **37,** 501–518.

Chandrasekharan, R., Moriarty, M. M., & Wright, G. P. (1994). Testing for unreliable estimators and insignificant forecasts in combined forecasts. *Journal of Forecasting,* **13,** 611–624.

Chhibber, S., & Apostolakis, G. (1993). Some approximations useful to the use of dependent information sources. *Reliability Engineering and System Safety,* **42,** 67–86.

Chhibber, S., Apostolakis, G., & Okrent, D. (1992). A taxonomy of the use of expert judgments in safety studies. *Reliability Engineering & System Safety,* **38,** 27–45.

Clemen, R. T. (1985). Extraneous expert information. *Journal of Forecasting,* **4,** 329–348.

Clemen, R. T. (1986). Calibration and the aggregation of probabilities. *Management Science,* **32,** 312–314.

Clemen, R. T. (1989). Combining forecasts: A review of annotated bibliography. *International Journal of Forecasting,* **5,** 559–583.

Clemen, R. T., Jones, S. K., & Winkler, R. L. (1996). Aggregating forecasts: An empirical evaluation of some Bayesian methods. In D. Berry, K. Chaloner, & J. Geweke (Eds.), *Bayesian Statistics and Econometrics: Essays in Honor of Arnold Zellner* (pp 3–13). New York: Wiley.

Clemen, R. T., & Murphy, A. H. (1986). Objective and subjective precipitation probability forecasts: Statistical analysis of some interrelationships. *Weather and Forecasting,* **1,** 56–65.

Clemen, R. T., & Reilly, T. (1999). Correlations and copulas for decision and risk analysis. *Management Science,* **45,** 208–224.

Clemen, R. T., & Winkler, R. L. (1985). Limits for the precision and value of information from dependent sources. *Operations Research,* **33,** 427–442.

Clemen, R. T., & Winkler, R. L. (1987). Calibrating and combining precipitation probability forecasts. In R. Viertl (Ed.), *Probability and Bayesian Statistics* (pp. 97–110). New York: Plenum.

Clemen, R. T., & Winkler, R. L. (1990). Unanimity and compromise among probability forecasters. *Management Science,* **36,** 767–779.

Clemen, R. T., & Winkler, R. L. (1993). Aggregating point estimates: A flexible modeling approach. *Management Science,* **39,** 501–515.

Cooke, R. M. (1991). *Experts in Uncertainty: Opinion and Subjective Probability in Science.* New York: Oxford University Press.

Dalkey, N. C. (1969). *The Delphi method: An experimental study of group opinions.* Report No. RM-5888-PR. The Rand Corporation.

Dalkey, N. C., & Brown, B. (1971). *Comparison of group judgment techniques with short-range predictions and almanac questions.* Report No. R-678-ARPA. The RAND Corporation.

Dall'Aglio, G., Kotz, S., & Salinetti, G. (1991). *Advances in Probability Distributions with Given Marginals: Beyond the Copulas.* Dordrecht, Netherlands: Kluwer.

Davis, J. (1992). Some compelling intuitions about group consensus decisions, theoretical and empirical research, and interpersonal aggregation phenomena: Selected examples, 1950–1990. *Organizational Behavior and Human Decision Processes,* **52,** 3–38.

Dawes, R. M., Faust, D., & Meehl, P. A. (1989). Clinical versus actuarial judgment. *Science,* **243,** 1668–1673.

de Finetti, B. (1937). La Prévision: Ses Lois Logiques, Ses Sources Subjectives. *Annales De L'Institut Henri Poincaré,* **7**.

Delbecq, A. L., Van de Ven, A. H., & Gustafson, D. H. (1975). *Group Techniques for Program Planning.* Glenview, IL: Scott Foresman.

Einhorn, H. J., Hogarth, R. M., & Klempner, E. (1977). Quality of group judgment. *Psychological Bulletin,* **84,** 158–172.

EPRI (1986). *Seismic Hazard Methodology for the Central and Eastern United States. Vol. 1: Methodology.* NP-4/26. Palo Alto, CA: Electric Power Research Institute.

Ferrell, W. R. (1985) Combining individual judgments. In G. Wright (Ed.), *Behavioral Decision Making* (pp. 111–145). New York: Plenum.

Fischer, G. (1975). *An experimental study of four procedures for aggregating subjective probability assessments.* Technical report 75-7. Decisions and Designs, Inc.

Flores, B. E., & White, E. M. (1989). Subjective vs. objective combining of forecasts: An experiment. *Journal of Forecasting,* **8,** 331–341.

French, S. (1981). Consensus of opinion. *European Journal of Operational Research,* **7,** 332–340.

French, S. (1985). Group consensus probability distributions: A critical survey. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian Statistics 2* (pp. 183–197). Amsterdam: North-Holland.

Gelfand, A. E., Mallick, B. K., & Dey, D. K. (1995). Modeling expert opinion rising as a partial probabilistic specification. *Journal of the Ameican Statistical Association,* **90,** 598–604.

Genest, C. (1984). Pooling operators with the marginalization property. *Canadian Journal of Statistics,* **12,** 153–163.

Genest, C., & McConway, K. J. (1990). Allocating the weights in the linear opinion pool. *Journal of Forecasting,* **9,** 53–73.

Genest, C., & Schervish, M. J. (1985). Modeling expert judgments for Bayesian updating. *Annals of Statistics,* **13,** 1198–1212.

Genest, C., & Zidek, J. V. (1986). Combining probability distributions. A Critique and annotated bibliography. *Statistical Science,* **1,** 114–148.

Gigone, D., & Hastie, R. (1997). Proper analysis of the accuracy of group judgments. *Psychological Bulletin,* **121,** 149–167.

Gokhale, D. V., & Press, S. J. (1982). Assessment of a prior distribution for the correlation coefficient in a bivariate normal distribution. *Journal of the Royal Statistical Society, Series A,* **145,** 237–249.

Goodman, B. (1972). Action selection and likelihood estimation by individuals and groups. *Organizational Behavior and Human Performance,* **7,** 121–141.

Gough, R. (1975). The effects of group format on aggregate subjective probability distributions. In *Utility, Probability, and Human Decision Making.* Dordrecht, Netherlands: Reidel.

Gustafson, D. H., Shukla, R. U., Delbecq, A., & Walster, G. W. (1973). A comparative study of differences in subjective likelihood estimates made by individuals, interacting groups, Delphi groups, and nominal groups. *Organizational Behavior and Human Performance,* **9,** 280–291.

Hastie, R. (1986). Review essay: Experimental evidence on group accuracy. In B. Grofman & G. Owen (Eds.), *Information Pooling and Group Decision Making: Proceedings of the Second University of California, Irvine, Conference on Political Economy.* Greenwich, CT: JAI Press.

Hammitt, J. K., & Shlyakhter, A. I. (1999). The expected value of information and the probability of surprise. *Risk Analysis,* **19,** 135–152.

Hill, G. W. (1982). Group vs. individual performance: Are N + 1 heads better than one? *Psychological Bulletin,* **91,** 517–539.

Hogarth, R. M. (1977). Methods for aggregating opinions. In H. Jungermann & G. DeZeeuw (Eds.), *Decision Making and Change in Human Affairs* (pp. 231–255). Dordrecht, Netherlands: Reidel.

Hogarth, R. M. (1987). *Judgment and Choice: 2nd Ed.* Chichester, England: Wiley.

Hora, S. C. (1992). Acquisition of expert judgment: Examples from risk assessment. *Journal of Energy Engineering,* **118,** 136–148.

Hora, S. C., Dodd, N. G., & Hora, J. A. (1993). The use of decomposition in probability assessments on continuous variables. *Journal of Behavioral Decision Making,* **6,** 133–147.

Hora, S. C., & Iman, R. L. (1989). Expert opinion in risk analysis: The NUREG-1150 methodology. *Nuclear Science and Engineering,* **102,** 323–331.

Innami, I. (1994). The quality of group decisions, group verbal behavior, and intervention. *Organizational Behavior and Human Decision Processes,* **60,** 409–430.

Janis, I. L. (1982). *Groupthink: Psychological Studies of Policy Decisions and Fiascoes, 2nd Ed.* Boston: Houghton Mifflin.

Janis, I. L., & Mann, L. (1977). *Decision Making.* New York: Free Press.

Jouini, M. N., & Clemen, R. T. (1996). Copula models for aggregating expert opinions. *Operations Research,* **44,** 444–457.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology,* **3,** 430–454.

Kaplan, S. (1990). 'Expert information' vs 'expert opinions': Another approach to the problem of eliciting/combining/using expert knowledge in PRA. *Journal of Reliability Engineering and System Safety, 39.*

Keeney, R. L., & von Winterfeldt, D. (1989). On the uses of expert judgment on complex technical problems. *IEEE Transactions on Engineering Management,* **36,** 83–86.

Keeney, R. L., & von Winterfeldt, D. (1991). Eliciting probabilities from experts in complex technical problems. *IEEE Transactions on Engineering Management,* **38,** 191–201.

Lawrence, M. J., Edmundson, R. H., & O'Connor, M. J. (1986). The accuracy of combining judgmental and statistical forecasts. *Management Science,* **32,** 1521–1532.

Lindley, D. V. (1983). Reconciliation of probability distributions. *Operations Research,* **31,** 866–880.

Lindley, D. V. (1985). Reconciliation of discrete probability distributions. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian Statistics 2* (pp. 375–390). Amsterdam: North-Holland.

Linstone, H. A., & Turoff, M. (1975). *The Delphi Method: Techniques and Applications.* Reading, MA: Addison-Wesley.

Lipscomb, J., Parmigiani, G., & Hasselblad, V. (1998). Combining expert judgment by hierarchical modeling: An application to physician staffing. *Management Science,* **44,** 149–161.

Lock, A. (1987). Integrating group judgments in subjective forecasts. In G. Wright & P. Ayton (Eds.), *Judgmental Forecasting* (pp. 109–127). Chichester, England: Wiley.

MacGregor, D., Lichtenstein, S., & Slovic, P. (1988). Structuring knowledge retrieval: An analysis of decomposing quantitative judgments. *Organizational Behavior and Human Decision Processes,* **42,** 303–323.

Makridakis, S., and Winkler, R. L. (1983). Averages of forecasts: Some empirical results. *Management Science,* **29,** 987–996.

Mendel, M. B., & Sheridan, T. B. (1989). Filtering information from human experts. *IEEE Transactions on Systems, Man, and Cybernetics,* **36,** 6–16.

Merkhofer, M. W. (1987). Quantifying judgmental uncertainty: Methodology, experience, and insights. *IEEE Transactions on Systems, Man, and Cybernetics,* **17,** 741–752.

Morgan, M. G., & Henrion, M. (1990). *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis.* Cambridge, MA: Cambridge University Press.

Morgan, M. G., & Keith, D. W. (1995). Subjective judgments by climate experts. *Environmental Science and Technology,* **29,** 468–476.

Morris, P. A. (1974). Decision analysis expert use. *Management Science,* **20,** 1233–1241.

Morris, P. A. (1977). Combining expert judgments: A Bayesian approach. *Management Science,* **23,** 679–693.

Morris, P. A. (1983). An axiomatic approach to expert resolution. *Management Science,* **29,** 24–32.

Mosleh, A., Bier, V. M., & Apostolakis, G. (1987). A critique of current practice for the use of expert opinions in probabilistic risk assessment. *Reliability Engineering and System Safety,* **20,** 63–85.

Myers, D. G., & Lamm, H. (1975). The polarizing effect of group discussion. *American Scientist,* **63,** 297–303.

Newbold, P., & Granger, C. W. J. (1974). Experience with forecasting univariate time series and the combination of forecasts. *Journal of the Royal Staistical Society, Series A,* **137,** 131–149.

Otway, H., & von Winterfeldt, D. (1992). Expert judgment in risk analysis and management: Process, context, and pitfalls. *Risk Analysis,* **12,** 83–93.

Parenté, F. J., & Anderson-Parenté, J. K. (1987). Delphi inquiry systems. In G. Wright & P. Ayton (Eds.), *Judgmental Forecasting* (pp. 129–156). Chichester, England: Wiley.

Park, W. (1990). A review of research on groupthink. *Journal of Behavioral Decision Making,* **3,** 229–245.

Phillips, L. D. (1984). A theory of requisite decision models. *Acta Psychologica,* **56,** 29–48.

Phillips, L. D. (1987). On the adequacy of judgmental forecasts. In G. Wright & P. Ayton (Eds.), *Judgmental Forecasting* (pp. 11–30). Chichester, England: Wiley.

Phillips, L. D., & Phillips, M. C. (1990). *Facilitated work groups: Theory and practice.* Unpublished manuscript, London School of Economics and Political Science.

Plous, S. (1993). *The Psychology of Judgment and Decision Making.* New York: McGraw-Hill.

Ravinder, H. V., Kleinmuntz, D. N., & Dyer, J. S. (1988). The reliability of subjective probabilities obtained through decomposition. *Management Science,* **34,** 186–199.

Reagan-Cirincione, P. (1994). Improving the accuracy of group judgment; A process intervention combining group facilitation, social judgment analysis, and information technology. *Organizational Behavior and Human Decision Processes,* **58,** 246–270.

Rohrbaugh, J. (1979). Improving the quality of group judgment: Social judgment analysis and the Delphi technique. *Organizational Behavior and Human Performance,* **24,** 73–92.

Savage, L. J. (1954). *The Foundations of Statistics.* New York: Wiley.

Schmittlein, D. C., Kim, J., & Morrison, D. G. (1990). Combining forecasts: Operational adjustments to theoretically optimal rules. *Management Science,* **36,** 1044–1056.

Seaver, D. A. (1978). *Assessing probability with multiple individuals: Group interaction versus mathematical aggregation* (Report No. 78-3). Social Science Research Institute, University of Southern California.

Shlyakhter, A. I. (1994). Improved framework for uncertainty analysis: Accounting for unsuspected errors. *Risk Analysis,* **14,** 441–447.

Shlyakhter, A. I., Kammen, D. M., Brodio, C. L., & Wilson, R. (1994). Quantifying the credibility of energy projections from trends in past data: The U.S. energy sector. *Energy Policy,* **22,** 119–130.

Sniezek, J. A. (1989). An examination of group process in judgmental forecasting. *International Journal of Forecasting,* **5,** 171–178.

Sniezek, J. (1992). Groups under uncertainty: An examination of confidence in group decision making. *Organizational Behavior and Human Decision Processes,* **52,** 124–155.

Sniezek, J. A., & Henry, R. A. (1989). Accuracy and confidence in group judgment. *Organizational Behavior and Human Decision Processes,* **43,** 1–28.

Sniezek, J. A., & Henry, R. A. (1990). Revision, weighting, and commitment in consensus group judgment. *Organizational Behavior and Human Decision Processes,* **45,** 66–84.

Staël von Holstein, C.-A. S. (1972). Probabilistic forecasting: An

experiment related to the stock market. *Organizational Behavior and Human Performance,* **8,** 139–158.
Stone, M. (1961). The opinion pool. *Annals of Mathematical Statistics,* **32,** 1339–1342.
Tindale, R. S., Sheffey, S., & Filkins, J. (1990). Conjunction errors by individuals and groups. Paper presented at the annual meeting of the Society for Judgment and Decision Making, New Orleans, LA.
Uecker, W. C. (1982). The quality of group performance in simplified information evaluation. *Journal of Accounting Research,* **20,** 388–402.
West, M. (1992). Modelling agent forecast distributions. *Journal of the Royal Statistical Society B,* **54,** 553–567.
West, M., & Crosse, J. (1992). Modelling probabilistic agent opinion. *Journal of the Royal Statistical Society B,* **54,** 285–299.
Winkler, R. L. (1968). The consensus of subjective probability distributions. *Management Science,* **15,** 361–375.
Winkler, R. L. (1981). Combining probability distributions from dependent information sources. *Management Science,* **27,** 479–488.
Winkler, R. L., & Clemen, R. T. (1992). Sensitivity of weights in combining forecasts. *Operations Research,* **40,** 609–614.
Winkler, R. L., & Makridakis, S. (1983). The combination of forecasts. *Journal of the Royal Statistical Society, Series A,* **146,** 150–157.
Winkler, R. L., & Poses, R. M. (1993). Evaluating and combining physicians' probabilities of survival in an intensive care unit. *Management Science,* **39,** 1526–1543.
Winkler, R. L., Wallsten, T. S., Whitfield, R. G., Richmond, H. M., Hayes, S. R., & Rosenbaum, A. S. (1995). An assessment of the risk of chronic lung injury attributable to long-term ozone exposure. *Operations Research,* **43,** 19–28.
Wright, G., Saunders, C., & Ayton, P. (1988). The consistency, coherence and calibration of holistic, decomposed, and recomposed judgmental probability forecasts. *Journal of Forecasting,* **7,** 185–199.
Wu, J. S., Apostolakis, G., & Okrent, D. (1990). Uncertainties in system analysis: Probabilistic versus nonprobabilistic theories. *Reliability Engineering & System Safety,* **30,** 163–181.